Visualizing Word Similarity Clusters

Thomas Wielfaert & Kris Heylen
Quantitative Lexicology and Variational Linguistics (QLVL), KU Leuven

Purpose: Visual analysis of the different meanings and uses of a polysemous word through interactive exploration of a large number of corpus concordances. These words occurrences are modelled through statistical Word Sense Disambiguation. Clustering prior to visualization allows to zoom in and out to different levels of detail.

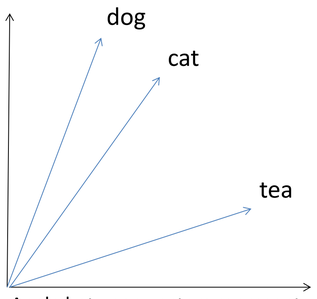
1. Word Spaces: The Word-Context Matrix

Word Spaces model meaning in terms of frequency distributions of words over co-occurring context words. The idea is based on the Distributional Hypothesis in Linguistics: Words that appear in the same context tend to have a similar meanings (Harris 1954).

Type-level Spaces (first-order co-occurrences):

Toy corpus:
1) The **dog** barked loud at the vet.
2) The **cat** scratched the vet.
3) Coffee tastes better than **tea**.

	bark	vet	scratch	taste	coffee
dog	**1*5.0**	**1*3.0**	0*.1.0	0*0.2	0*0.1
cat	0*0.1	**1*4.0**	**1*6.0**	0*0.3	0*0.1
tea	0*0.1	0*0.3	0*0.1	**1*4.0**	**1*2.5**



Angle between vectors represents cosine similarity

Optional weighting:
Not every context word is equally informative for the meaning of the target. Therefore, the raw frequencies are multiplied by a weight which represents the collocational strength between target in context.

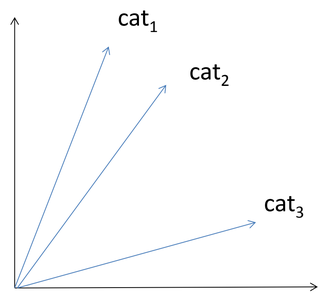
Literature: see Turney and Pantel (2010) for an overview of the types of Semantic Vector Spaces.

Token-level Spaces (second-order co-occurrences):

Toy corpus of *cat*:
1) Blofeld was stroking the purring **cat**$_1$ on his lap.
2) The black dog barked at the **cat**$_2$ in the tree.
3) The cadet was sailing his **cat**$_3$ against the wind.

	...	cup	blow	fur	pet	sea	...
lap		1.2	0.2	2.4	4.1	0.1	
		+	+	+	+	+	
purr		0.1	0.6	4.1	4.6	0.0	
		+	+	+	+	+	
stroke		0.4	1.2	3.2	4.5	0.2	
SUM		1.7	2.0	9.7	13.2	0.3	
		÷3	÷3	÷3	÷3	÷3	
AVERAGE		0.6	0.7	3.2	4.4	0.1	

Second-order co-occurrences for cat$_1$



	...	cup	blow	fur	pet	sea	...
Blofeld was **stroking** the **purring** CAT in his **lap**		0.6	0.7	3.2	4.4	0.1	
The black **dog barked** at the **CAT** in the **tree**		0.2	1.5	2.5	3.4	0.8	
The **cadet** was **sailing** his **CAT** against the **wind**		0.3	3.7	0.2	0.3	3.7	

Token vectors of second order co-occurrences.

	Blofeld was **stroking** the **purring** CAT in his **lap**	The black **dog barked** at the **CAT** in the **tree**	The **cadet** was **sailing** his **CAT** against the **wind**
Blofeld was **stroking** the **purring** CAT in his **lap**	1	0.96	0.18
The black **dog barked** at the **CAT** in the **tree**		1	0.42
The **cadet** was **sailing** his **CAT** against the **wind**			1

Token by token similarity matrix.

2. Visualizing token space

Dimension reduction
Word space is reduced to 2 dimensions through standard technique for visualizing high-dimensional similarity matrices; Multidimensional Scaling (MDS).

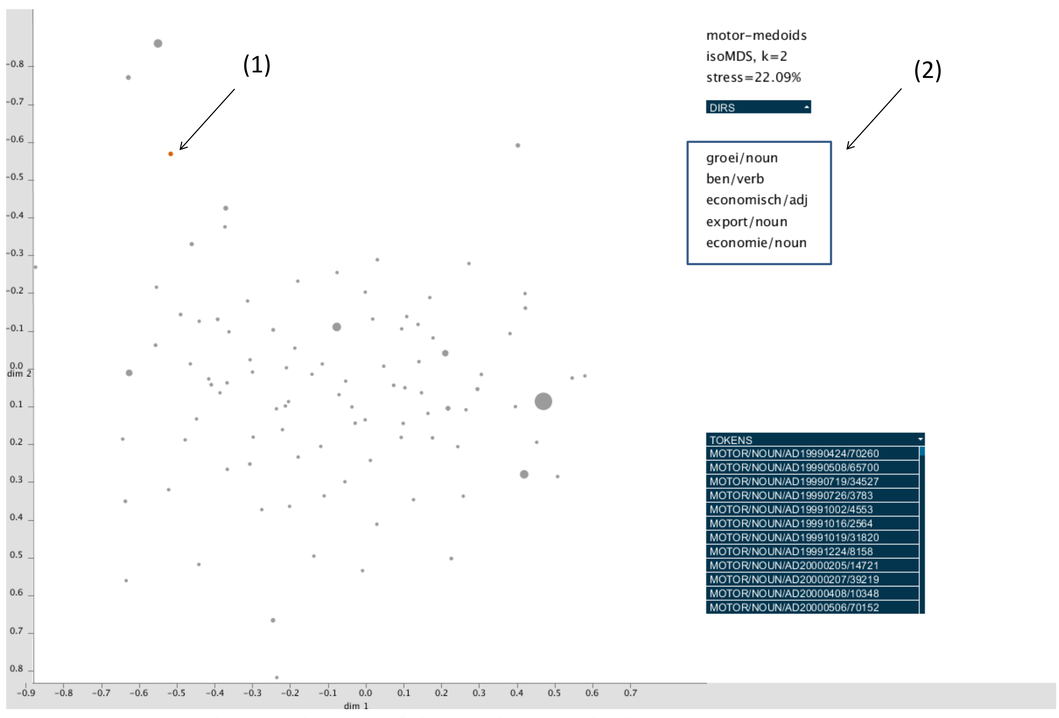
Corpus
Twente Nieuws Corpus (TwNC): ± 500 M words from Dutch newspapers (Ordelman 2002)

Case study
Meanings for the polysemous Dutch word *motor* according to the online dictionary Algemeen Nederlandse Woordenboek (ANW)*:*
motor$_1$*: engine*
motor$_2$*: motorcycle*

Visualization principles
The number of tokens that can be displayed in an informative way, i.e. without disturbing overlap of data points, is limited to 200-300 tokens per plot. This visualization has been implemented according to Shneiderman's Visual Information-Seeking Mantra (1996): overview first, zoom and filter, then details-on-demand.

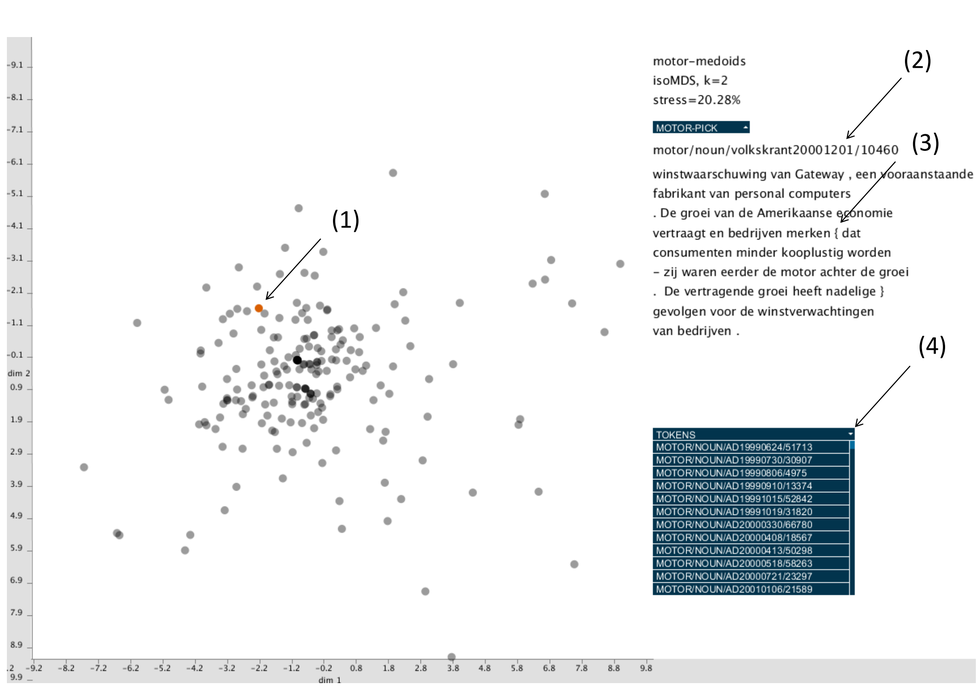
3. k-medoids clustering

Clustering a large number of tokens (n=12690) by using a k-medoids clustering algorithm (k=100) to get an interpretable structure.



Overview scatter plot visualization of the resulting medoids.
(1) Selected cluster.
(2) Top-5 context words in the cluster, in casu: *groei* (growth), *ben* (first person of to be), *economisch* (economical), *export* and *economie*.
Note: Dot size corresponds to cluster size

4. Zoom on *motor* cluster



Details
Visualization of the cluster selected in the plot above.
(1) One token selected.
(2) Token label structured as following: word type/POS/newspaper+date/line number.
(3) Context window around target word, here in bold:

De groei van de Amerikaanse economie vertraagt en bedrijven merken ***dat consumenten minder kooplustig worden – zij waren eerder de motor achter de groei. De vertragende groei heeft nadelige*** *gevolgen voor de winstverwachtingen van de bedrijven*

The growth of the American economy slows down and companies notice **that consumers less eager to buy become – they were earlier the engine behind the growth. The slowing growth has adverse** consequences for the profit expectations of companies.

(4) Drop-down menu to select a token by label.

Metaphorical extension
The token shown here is a metaphorical extension of **motor**$_1$. It refers to the thriving force behind the economy, which seems to be the core sense of *motor* represented in this cluster.

5. References

Harris, Z. (1954). Distributional structure. *Word*, 10(23): 146-162.
Ordelman, R.J.F., Twente nieuws corpus (TwNC). Technical report, August 2002.
Shneiderman, B, The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 336-343, Washington. IEEE Computer Society Press, 1996.
Turney, P.D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1), 141-188.

Thomas Wielfaert, Kris Heylen thomas.wielfaert@kuleuven.be kris.heylen@kuleuven.be
University of Leuven/Faculty of Arts/Quantitative Lexicology and Variational Linguistics